Call sign intelligibility improvement using a spatial auditory display

N94-33617

Durand R. Begault, Ph.D.
Human Interface Research Branch
Aerospace Human Factors Research Division
NASA - Ames Research Center
MS 262-2, Moffett Field, CA 94035-1000

Abstract

A spatial auditory display was designed for separating the multiple communication channels usually heard over one ear to different virtual auditory positions. The single 19" rack mount device utilizes digital filtering algorithms to separate up to four communication channels. The filters use four different binaural transfer functions, synthesized from actual outer ear measurements, to impose localization cues on the incoming sound. Hardware design features include "fail-safe" operation in the case of power loss, and microphone/headset interfaces to the mobile launch communication system in use at NASA Kennedy Space Center. An experiment designed to verify the intelligibility advantage of the display used 130 different call signs taken from the communications protocol used at NASA KSC. A 6 to 7 dB intelligibility advantage was found when multiple channels were spatially displayed, compared to monaural listening. The findings suggest that the use of a spatial auditory display could enhance both occupational and operational safety and efficiency of NASA operations. (Supported by NASA Ames and NASA KSC Director's Discretionary Funding).

1. INTRODUCTION

1.1 Application to NASA communication systems.

During fiscal year 1992, NASA Director's Discretionary Funding was received from Ames Research Center (ARC) and John F. Kennedy Space Center (KSC) by Drs. E. M. Wenzel and D. R. Begault, to develop a four channel spatial auditory display for application to multiple channel speech communication systems in use at KSC. A previously specified design (Begault & Wenzel, 1990; Begault, 1992a) was used to fabricate a prototype device, which was completed in February, 1993.[1] This prototype places four different communication channels in virtual auditory positions about the listener, by digitally filtering each input channel with binaural head-related transfer function (HRTF) data. Listening over headphones, one has a spatial sense of each channel originating from a unique position outside the head; i.e., as if four people were standing about you, speaking from different directions.

Input channels to the spatial auditory display can be assigned to any position because the design uses four removable EPROMs [2], with each EPROM corresponding to a particular target position. The EPROMs themselves can contain a binaural HRTF for any given position and measured ear. Hence, an important research question is to determine which four positions would be optimal for speech intelligibility of multiple sound sources. To begin to answer this question, the current investigation focused on what single spatialized azimuth position yielded maximal intelligibility against noise. This was accomplished by measuring intelligibility thresholds at 30° azimuth increments. Intelligibility is defined here as correct identification of a spatialized call sign (signal) against diotic [3] speech babble (noise).

The KSC communications handbook (NASA-KSC, 1991) indicates a list of over 3000 call signs, most of which are spoken as four individual letters-- e.g., "NTOC". Communication personnel who monitor multiple radio frequencies must be able to hear these four letters clearly against speech. The use of speech babble as a noise source has been used in several studies investigating binaural hearing for

---

[1] Tom Erbe (Mills College, Sterling Software) implemented the firmware and hardware design into the prototype.

[2] EPROM = erasable-programmable-read-only memory chip.

[3] "Diotic" playback is defined as a single audio channel presented to both ears.

Presented at 1993 SOAR (Space Operations, Applications and Research) Conference, NASA Johnson Space Center, Houston

communication systems contexts (e.g., Pollack, & Pickett 1958). This study concludes with a first approximation of the answer to what HRTF positions are best used in the filter EPROMs within the prototype.

1.2 Binaural advantages and speech intelligibility.

The relationship between binaural hearing and the development of improved communication systems has been understood for over 45 years (Licklider, 1948; see reviews in Blauert, 1983; Zurek, 1993). As opposed to monotic (one ear) listening-- the typical situation in communications operations-- binaural listening allows a listener to use head-shadow and binaural interaction advantages simultaneously (Zurek, 1993). The head-shadow advantage is an acoustical phenomenon, caused by the interaural level differences that occur when a sound moves closer to one ear relative to the other. Because of the diffraction of lower frequencies around the head from the near ear to the far ear, only frequencies above approximately 1.5 kHz are shadowed in this way. The binaural interaction advantage is a psychoacoustic phenomenon due to the auditory system's comparison of binaurally-received signals (Levitt & Rabiner, 1967; Zurek, 1993).

Many studies have focused on binaural advantages for both for detecting a signal against noise (the binaural masking level difference, or BMLD), and for improving speech intelligibility (the binaural intelligibility level difference, or BILD). Studies of BMLDs and BILDs involve manipulation of signal processing variables affecting either signal, noise, or both. The manipulation can involve phase inversion, time delay, and/or filtering.

Recently, speech intelligibility studies by Bronkhorst and Plomp (1988; 1992) have used a mannequin head to impose the filtering effects of the HRTF on both signal and noise sources. The HRTFs were used in either an unaltered condition, or with either time or amplitude components removed. Their results, summarized in Figure 1, show a 6 to 10 dB advantage with the signal at 0° azimuth and speech-spectrum noise moved off axis, compared to the condition where speech and noise originated from the same position. Figure 1 also shows lower BILDs when either interaural time or amplitude differences are removed from the stimuli. This suggested the inclusion of HRTF filtering within a binaural display for speech communication systems (ref. Begault & Wenzel, 1990; Begault & Wenzel, 1992). According to a model proposed by Zurek (1993), based on averaged HRTFs specified in Shaw & Vaillancourt (1985), the average binaural advantage (speech signal fixed at 0°, noise uniformly distributed across all azimuths, head free to move) is around 5 dB, with head shadowing contributing about 3 dB and binaural-interaction about 2 dB.

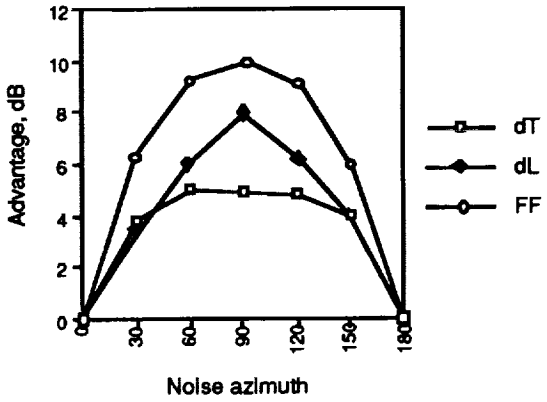

Figure 1. *Data from Bronkhorst and Plomp (1988) for speech intelligibility gain. All stimuli were recorded with a mannequin head. Speech signal fixed at 0°; noise moved along azimuth at 0° elevation. FF= data including effects of the HRTF; dT = same data with binaural amplitude differences removed; dL = same data but with binaural time differences removed.*

Another advantage for binaural speech reception relates to the ability to switch voluntarily between multiple channels, or "streams", of information (Bregman, 1990; Deutsch, 1983). The improvement in the detection of a desired speech signal against multiple speakers commonly referred to as the "cocktail party effect" (Cherry, 1953; Cherry & Taylor, 1954) is explained by Bregman (1990) as a form of auditory stream segregation. This situation was found to parallel the multiple channel listening requirements of communication personnel, such as test directors (NTDs) at KSC.

2. METHOD

2.1 Stimuli.

The signal portion of the stimulus was drawn from a list of 130 four letter call signs, selected from the KSC communication handbook (NASA-KSC, 1991). The 130 call signs used in the experiment were selected randomly so that groups of five began with a unique letter of the alphabet. A single male voice was used, with each letter of the call sign spoken discontinuously over a duration of about two seconds. Recordings took place in sound-proof booth, using an AKG C451-EB microphone at a distance of 6 inches. Once digitized, each call sign combination was normalized in amplitude, and then scaled to have equal long-term r.m.s. measurement values.

The speech babble used for the noise portion of the stimulus consisted of multiple layers of voices: two layers were from different airport control tower frequencies, containing both female and male voices, with silent intervals of more than .2 seconds deleted; and two recordings of different male voices reading technical repair manuals, one played backwards, the other pitch shifted upwards 4 semitones. The result was a dense speech layer in which words could occasionally be distinguished, but semantic content was lost.

The noise and speech were digitally stored as separate channels of stereo sound files (see Figure 2), using an Apple Macintosh II fx and Digidesign's ProTool hardware and software. The duration of each sound file used in each stimulus presentation was adjusted to 5 seconds, with the noise channel faded in and out over the first and last 0.5 seconds. The signal was always presented 1.5 seconds into the sound file, allowing subjects to predict its onset.

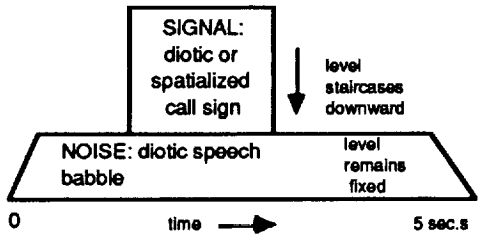

Figure 2. *Stimulus soundfile arrangement*

Each of the 130 separate noise-signal sound files was played through signal processing software and hardware, using a Crystal River Engineering Convolvotron that also served as the experimental software host computer (see Wenzel, 1992, for additional information on the hardware). Upon playback, the Convolvotron passed the speech babble channel unaltered to both ears. Mixed in with this noise was the two-channel signal, after software intensity scaling and HRTF-based spatialization to azimuths at 30 degree increments between 30° - 330° (all at 0° elevation). A diotic control condition was also used for the signal, where the spatialization was bypassed and only intensity scaling was used.

The minimum-phase HRTFs used for the spatialization were reconstructed from actual HRTF measurements as described in Kistler & Wightman (1992). The original measurements used were of one subject (SDO in Wightman & Kistler, 1989), with the headphone frequency response (Sennheiser HD-430) divided out of the HRTF. Although the same model of headphone was used for the subjects in this experiment, non-linearities in reproducing the HRTF were introduced as a result of the interaction between different pinnae and the headphone chambers. Data on localization error of speech with non-individualized HRTFs can be found in Begault & Wenzel (1991) and Begault (1992b).

2.2 Subjects

Five subjects (4 males, 1 female), were paid $5.59 an hour to participate in the study over two three hour sessions. This was the "naive subjects" group in that they had no exposure to the call sign list. Another group of 3 lab personnel (3 males) who had previous exposure to the call sign list constituted the "experienced subjects" group; their data is analyzed separately from the naive subject group. This group included a subject whose voice was used in the signal.

All subjects were evaluated for normal hearing from 0.1 - 8 kHz in a pure tone audiometer test. Subjects were given a training session before starting the experiment to familiarize themselves with the computer, the time when to expect the signal in relation to the noise, and the

procedure for entering responses. This training session consisted of a dummy block where the level of the signal was clearly audible against the noise, and was never scaled. The formal blocks were begun after approximately 20 trials.

2.3 Procedure

Software was developed by Phil Stone (Sterling Software) for presenting stimuli and gathering data from subjects using an interleaved, transformed up-down "staircase" method (Levitt, 1970). The software varied the level of the signal against the noise, starting with a maximum stepsize interval of 6 dB, and decreasing to a minimum stepsize of 1 dB. The response sequences were evaluated in such a way as to determine the threshold at a 70.7% probability level (a "2 up, 1 down" procedure).

The decibel level between the diotic stimuli and the spatialized stimuli were considered to be equal with reference to the long-term r.m.s. value of speech-spectrum noise filtered by a left ear 0° HRTF (obtained from the same HRTF set used for the other spatialized positions). The playback level was around 55 dB SPL, when the noise and 0° HRTF-filtered calibration signals were played simultaneously.

Six blocks were administered to each subject over three or four days, with each block containing four staircases randomly chosen from the 11 possible spatial positions or the one diotic signal condition. The four staircases within each block were presented randomly, as were the 130 call sign-speech babble sound files used for a particular stimulus block. The staircases within the blocks were arranged so that ten threshold values were obtained from each subject for each spatial condition, and the diotic condition. No block contained two simultaneous staircases for the same spatial condition of the signal.

Upon hearing the stimulus, the subject typed the four letters they thought they had heard onto a computer keyboard, and then after a short pause the software would present the next trial. The duration to complete each block of four staircases was about 15 - 20 minutes. Testing was administered in a sound-proof booth. No feedback was given as to the correct identification of the call signs; the subjects were only notified when the 20 staircases within a particular block (4 spatial conditions times 5 staircases) were completed.

3. RESULTS

Figure 3 summarizes the data for the six naive subjects, and Figure 4 summarizes the data for the three experienced subjects. The mean values for each position were obtained before grouping the data by first subtracting each individual subject's threshold for the diotic signal vs. diotic speech babble condition. The results in Figures 3-4 show a greater intelligibility advantage as the signal is moved from to either side of the head; the advantage is maximal between 60° - 90° and 270° - 300°. These are locations where both head-shadowing is maximized, and where the binaural interaction advantage mechanism is given maximal time differences.

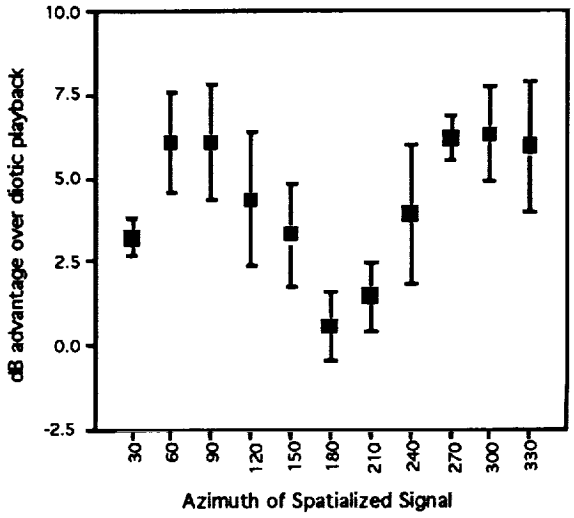

Figure 3. *Data for the naive subject group (4 males, 1 female). The mean value for the diotic signal condition were subtracted from each spatialized signal value. Standard deviation bars were based on the 10 staircase solutions obtained for each condition.*

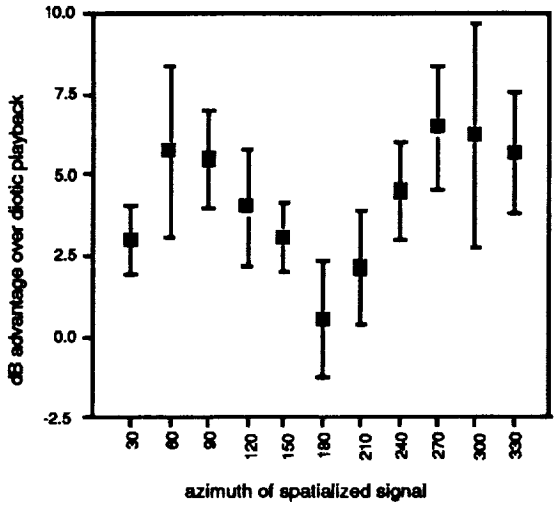

Figure 4. *Data for the experienced subject group (see Figure 3).*

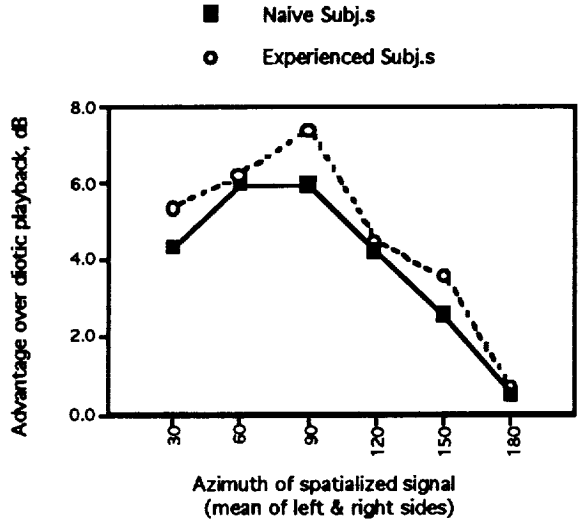

Figure 5. *Mean values from Figures 3-4 collapsed about symmetrical left-right positions.*

Figure 5 summarizes Figures 3-4, by showing the mean values for symmetrical left-right positions about the head. This suggests that, without reference to which side a sound is spatialized, the preferred order for HRTF-processing for maximal intelligibility is 60° or 90°, then 120°, then 30°, then 150°, and finally 180°. The latter is hardly better than performance with the diotic stimuli. Figure 5 also shows that the three experienced subjects achieved about a 1 dB additional intelligibility advantage over the five naive subjects. However, an analysis of variance revealed that no significant difference existed between these two subject categories, E (1,6) = 2.90, p = 0.14.

The mean values for four of the naive subjects had a pattern that followed the symmetrical trend of the overall mean shown in Figure 3; there seemed to be no preferred side to hear the signal. Contrasting this, the responses of one of the naive subjects had an asymmetrical trend, favoring right side positions over left side positions. This trend was similar to a potential subject whose data was excluded from the subject pool and the analysis above due to hearing loss at the left ear (between 20 - 35 dB HL at 4, 6, 8 and 12 kHz).

Figure 6 shows the results for these two subjects, along with the overall means from the naive subject group. Except for the 60° azimuth position, both of these subjects had a smaller advantage for left side positions compared to the overall mean, and right side positions show a greater advantage. Additional data would be needed to determine if there was a significant effect due to handedness or other factors (Deutsch, 1983). Nevertheless, a person with asymmetrical hearing loss similar to that experienced by the subject shown in Figure 6 could still benefit from using a 3-D auditory display. Gabriel, Koehnke and Colburn (1991) and Perrott, Sadralodabi, Saberi and Strybel (1991) have pointed out that, excluding severe hearing loss, no apparent relation between audiometric measurements and binaural performance can be established.

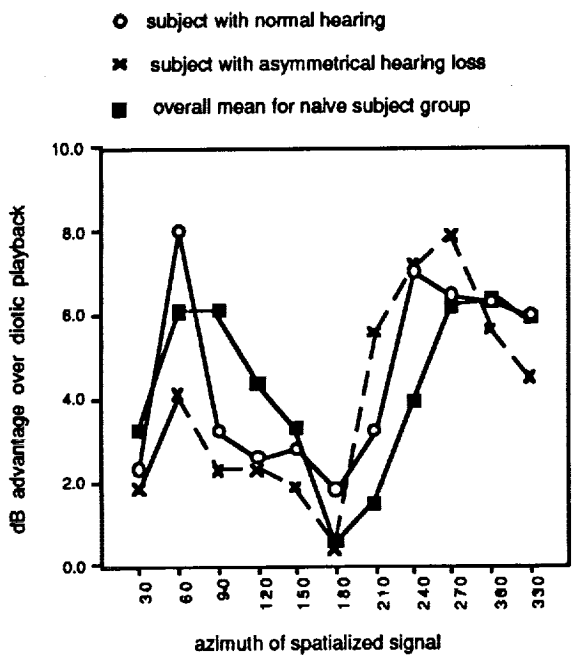

Figure 6. *Two subjects (one from the naive group, one subject w/ asymmetrical hearing loss) who tended to favor the right side positions over the left. Overall means (from Figure 3) shown for comparison.*

4. DISCUSSION

Overall, a 6-7 dB advantage for left and right 60° and 90° positions was found in the present study, which exceeds the binaural advantage cited in Zurek's model (1993) by 1-2 dB. This means that headphone listening with static spatial positions through the hardware prototype is as least as good as a normal hearing, binaural listener who is free to move their head. Although Bronkhorst and Plomp (1988) found a 10 dB advantage for a signal at 0° azimuth and speech-spectrum noise at 90°, their results are not directly comparable to those found here since both signal and noise were HRTF-filtered by their mannequin head, and in the present study the noise portion of the stimulus was diotic. The additional release from masking they found may have been attained through either HRTF-filtering of both signal and noise, the use of noise rather than speech babble, or both.

The results found here are limited by the fact that only one male speaker was used for the signal portion of the stimulus. In spite of the care taken in preparing the stimulus through digital editing, there is the potential that extraneous variation was introduced into the results because of the variability of spoken intelligibility (ANSI, 1989). Furthermore, the average spectrum of this particular speaker might have interacted differently with the HRTF filtering than that of another speaker (e.g., a female voice). Finally, the variability in HRTF measurements from different persons or reconstruction techniques could influence the results of any experiment that uses only one set of HRTFs. This is one reason the prototype was designed to allow interchangeable EPROMs- individuals could tailor systems to their best advantage by using a preferred set of HRTFs.

5. CONCLUSION

The advantage of a binaural auditory display for multiple communication channels has been demonstrated, through a case study of a single signal at incremented 30° azimuth positions against a diotic, speech babble noise source. The 6 -7 dB advantage for 60° and 90° HRTF-filtered speech represents a halving of the intensity (acoustic power) necessary for correctly identifying a four letter call signs typical of those used in communication systems at KSC. This reduction in the likelihood of misinterpreting call signs over communication systems is an important safety improvement for "high stress", human-machine interface contexts. The binaural advantage could also benefit communications personnel because the overall intensity of communications hardware could be reduced without sacrificing intelligibility. Lower listening levels over headphones could possibly reduce the risk of threshold shifts, the Lombard Reflex (raising the intensity of one's own voice; see Junqua, 1993), and overall fatigue, thereby making additional contributions to safety.

Overall, the findings here suggest that the use of a spatial auditory display could enhance both occupational and operational safety and efficiency of NASA operations. Additional studies are underway at Ames to simulate other applications scenarios within speech intelligibility experiments to determine the additional benefits, if any, of spatial audio communications displays.

6. ACKNOWLEDGMENTS

The assistance of Dr. Elizabeth M. Wenzel in all aspects of this work is gratefully

acknowledged. Further acknowledgment is due to Drs. Cynthia Null and Key Dismukes for comments on the manuscript; to Bill Williams and Jim Devault at KSC for their initiation of this project; and to Phil Stone, Rick Shrum and Tom Erbe for assistance in programming, experimental design, and prototype development.

7. BIBLIOGRAPHY

ANSI (1989). American National Standard: Method for measuring the intelligibility of speech over communication systems No. S3.2-1989. American National Standards Institute.

Begault, D. R. (1992a). Audio Spatialization Device for Radio Communications (Patent Disclosure No. ARC 12013-1CU). NASA-Ames Research Center.

Begault, D. R. (1992b). Perceptual effects of synthetic reverberation on three-dimensional audio systems. Journal of the Audio Engineering Society, 40(11), 895-904.

Begault, D. R., & Wenzel, E. M. (1990). Technical aspects of a demonstration tape for three-dimensional auditory displays (Technical Memorandum No. TM 102286). NASA-Ames Research Center.

Begault, D. R., & Wenzel, E. M. (1991). Headphone Localization of Speech Stimuli. In Proceedings of the Human Factors Society 35th Convention, (pp. 82-86). San Francisco: Santa Monica: Human Factors Society.

Begault, D. R., & Wenzel, E. M. (1992). Techniques and applications for binaural sound manipulation in human-machine interfaces. International Journal of Aviation Psychology, 2(1), 1-22.

Blauert, J. (1983). Spatial hearing: The psychophysics of human sound localization (J. Allen, Trans.). Cambridge: MIT Press.

Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Mass.: MIT Press.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. Journal of the Acoustical Society of America, 83(4), 1508-1516.

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. Journal of the Acoustical Society of America, 92(6), 3132-3139.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and two ears. Journal of the Acoustical Society of America, 25(5), 975 - 979.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments on the recognition of speech with one and with two ears. Journal of the Acoustical Society of America, 26, 549 - 554.

Deutsch, D. (1983). Auditory illusions, handedness, and the spatial environment. Journal of the Audio Engineering Society, 31(9), 607 - 623.

Gabriel, K. J., Koehnke, J., and Colburn, H. S. (1991). Frequency dependence of binaural performance in listeners with impaired binaural hearing. Journal of the Acoustical Society of America, 91, 336-347.

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. Journal of the Acoustical Society of America, 93(1), 510-524.

Kistler, D. J. , & Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. Journal of the Acoustical Society of America, 91(3), 1637-1647.

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49(2), 467-477.

Levitt, H., & Rabiner, L. R. (1967). Predicting binaural gain in intelligibility and release from masking of speech. Journal of the Acoustical Society of America, 42(4), 820-829.

Licklider, J. C. R. (1948). The influence of interaural phase relations upon the masking

of speech by white noise. Journal of the Acoustical Society of America, 20(150-159).

NASA-KSC (1991). KSC Operational Intercommunications System Call Sign/Word Handbook No. DRD 016, revision 18. John F. Kennedy Space Center.

Perrott, D. R., Sadralodabai, T., Saberi, K., & Strybel, T. Z. (1991). Aurally aided visual search in the central visual field: effects of visual load and visual enhancement of the target. Human Factors, 33(4), 389-400.

Pollack, I., & Pickett, J. M. (1958). Stereophonic listening and speech intelligbility. Journal of the Acoustical Society of America, 30(1), 131-133.

Shaw, E. A. G., & Vaillancourt, M. M. (1985). Transformation of sound-pressure level from the free field to the eardrum presented in numerical form. Journal of the Acoustical Society of America, 78(3), 1120-1122.

Wenzel, E. M. (1992). Localization in virtual acoustic displays. Presence: Teleoperators and Virtual Environments, 1, 80-107.

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. I : Stimulus synthesis. Journal of the Acoustical Society of America, 85(2), 858-867.

Zurek, P. M. (1993). Binaural Advantages and Directional Effects in Speech Intelligibility. In G. A. Studebaker & I. Hochberg (Eds.), Acoustical Factors Affecting Hearing Aid Performance Needham Heights, Mass.: Allyn and Bacon.

This article is available as NASA Technical Memorandum No. 104014. A demonstration tape of virtual acoustic displays is available from the author.